Plánfelismerés filmekben arcdetektálás segítségével

Marosvári Borbála
BME VIK Távközlési és Médiainformatikai Tanszék

Napjainkban a multimédiás tartalmak száma – a filmeket is ide értve – rendkívül gyorsan növekszik. Ezek rendszerezése, elemzése, a bennük történő keresés nagy kihívást jelent. Sok felcímkézett, annotált anyagra lenne szükség. Munkám során filmek automatikus annotációjával foglalkozom. Egy olyan rendszer, amely képes filmeket automatikusan felcímkézni, nagyban elősegíti a filmek szemantikus elemzését.

Filmelemzést sok szempont alapján végezhetünk: kameramozgás, kompozíció, mélységélesség, telítettség, egyensúly stb. A filmhez tartozó hangot is vizsgálhatjuk; beszélhetünk az éppen hallható hang vagy zörej redundanciájáról, narráció jelenlétéről, a zene diegetikusságáról, a kísérő zene aláfestő vagy ellenpontozó jellegéről stb. Ebben a munkámban egy olyan elkészített rendszert mutatok be, amely képes egy filmet annotálni a benne található plánokkal. A rendszer elkészítéséhez több programot, programkönyvtárat használtam: *ffmpeg*et a jelenethatárok megtalálásához, *OpenCV*-t a jellemzőkinyerő alkalmazáshoz, melyet C++ nyelven írtam meg, *Excel VBA*-t az elkészült adatsorok és manuális annotációk összeillesztésére, illetve *RapidMiner*t az osztályozáshoz és kiértékeléshez.

Plánok jellemzői

A film-, a TV- és a videóiparban a mozgóképnek azt a tulajdonságát, amely a szereplők és a környezetük egymáshoz való viszonyát, elhelyezkedését, illetve mozgását fejezi ki, plánnak nevezzük. A pontos terminológia a gyártási közeg függvényében változhat, de az alapvető jelentések megegyeznek.

Egy jeleneten vagy snitten belül többféle plán is lehet. Snittnek nevezzük a filmeknek azon szakaszait, melyeket vágások határolnak; azaz két vágás közötti folyamatos rész egy snitt; hossza bármilyen lehet. Ezáltal két különböző plán között az átmenet lehet folyamatos, vagy

vágás esetén ugrásszerű is. Az alábbiakban azokat a plánokat részletezem, amelyek munkám során az elérhető manuális annotációk alapján megkülönböztethetők voltak. Az illusztrációk az *Il deserto rosso* c. filmből vett képkockák.

Nagytotál plán: A hangsúly a környezet bemutatásán van, a szereplők, ha láthatók egyáltalán, akkor 30 méternél messzebb látszanak („Plán" 2015), Gyakran a jelenetek kezdő képkockái ilyenek, bemutatják a történés helyét, környezetét („Shot Types" 2012).

1. ábra: Nagytotál plán

Totál plán: A szereplők már jobban láthatók, de a hangsúly még mindig a környezetükön van. A példaképeken látható plánokon látható, hogy a nagytotál és a totál akár pontosan ugyanolyan távolságot is jelenthet, a két kép között balra svenkelés történt, a különbséget a kompozícióban kereshetjük.

2. ábra*:* Totál plán

Kistotál plán: A szereplők vagy egy nagyobb tárgy teljes egészében láthatók, kitöltik a képet, amennyire lehetséges, 30 méternél közelebbről látszanak („Plán" 2015).

3. ábra: Kistotál plán

Amerikai plán: a szekond plán egy variációja („Shot (filmmaking)" 2015), a szereplőket térdtől felfelé mutatja. Más néven cowboy plán, a westernekből eredő elnevezés, lényege, hogy a szereplő oldalán függő fegyver benne legyen a képben.

4. ábra: Amerikai plán

Szekond plán: egy vagy több szereplő fejét és testének egy részét, illetve valamilyen közepes méretű dolgot mutat meg. A valóságban is hasonló szögben és méretarányban látunk („Plán" 2015).

5. ábra: Szekond plán

Premier plán: egy egész fejet, testrészt vagy valamilyen kisebb méretű dolgot jelenít meg.

6. ábra: Premier plán

Szuper közeli plán: a premier plánnál még közelebbi állás, a szereplő fejének csak egy részlete látszik.

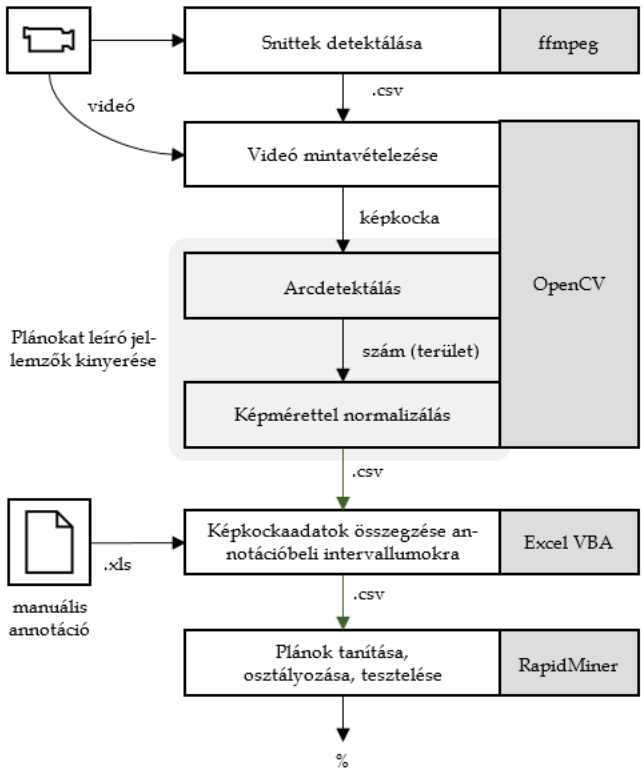
7. ábra: Szuper közeli plán

Ansnitt: Egy szereplő válla fölött, a szereplő mögül láthatjuk a kép tárgyát. Általában arra használják, hogy egy beszélgetést az első szereplő szemszögéből mutassanak („Shot Types" 2012).

8. ábra: Ansnitt

A rendszer felépítése

A cél egy olyan rendszer megalkotása volt, amellyel a rendszer bemenetére adott tetszőleges filmre a kimenetén megjelenik a film annotációja: a plánok megnevezése, időbélyegekkel ellátva. Munkám során a rendszert a multimédiaelemző rendszerek általános modellje alapján készítettem el, a nyers videóból kinyert különféle jellemzők alapján egy betanított tanuló algoritmussal sorolja be az alkalmazás a jeleneteket az előbb tárgyalt osztályokba. Általános felépítése az alábbi ábrán látható.



9. ábra: A rendszer általános blokkvázlata

Videó vizsgálata, mintavételezése

A vizsgált filmen első lépésként megkeresem, hogy hol van jelenetváltás, azaz vágás. Ehhez *ffmpeg*et használtam. A jelenetdetektor jelenetváltási érzékenységét a javasolt intervallum alapján (0,3–0,4) 0,35-ra állítottam, amely a próbavideók során megfelelőnek bizonyult. Az *ffmpeg* kimenetként

egy *.csv* fájlt állít elő, amely tartalmazza a talált jelenetek első képkockájának számát, időbélyegét ezredmásodpercben, a kép típusát (I, B, vagy P kép), illetve a döntés konfidenciáját.

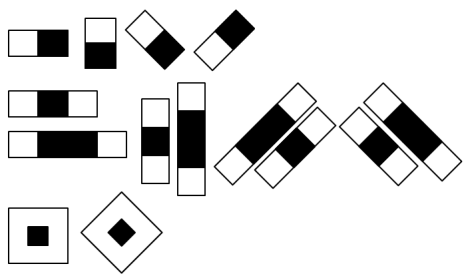
A videó mintavételezése konstans, 6 képkockánként mintavételezi a videót az alkalmazás, azonban a detektált jelenet első és utolsó képkockáját mindig megvizsgálja. Ebből kifolyólag egy-egy másodperc alatt eltérő számú képkockát mintavételezhet az alkalmazás. A képkockákhoz tartozó eredményeket a jellemzőkinyerés után végzi el a rendszer, így lehetőség van többféle csoportosításra is: másodpercenként, jelenetenként, vagy az annotációban meghatározott intervallumokként is megkaphatjuk a megállapított plánokat.

Jellemzőkinyerés

A plánok felismeréséhez a videón látható emberi arcok méretét használtam ki. Az alkalmazás minden vizsgált képkockán Viola-Jones arcdetektorral (Viola/Jones, 2001) keresi meg a látható arcokat. A Viola-Jones detektor egy *Haar wavelet*en alapuló kaszkádolt osztályozó *AdaBoost* tanulóalgoritmussal. A detektor többféle tárgy felismerésére is betanítható, de főként arcdetektálásra használják. A detektor három fő építőeleme a következő:

- integrálkép létrehozása;
- osztályozók tanítása *AdaBoost* alapú tanulóalgoritmussal;
- az osztályozók kaszkád struktúrába építése.

Az algoritmus futása során nem közvetlenül a vizsgált kép pixeleit értékeli ki, hanem a képi struktúrákat jól leíró *Haar wavelet*eket használ, amivel a feldolgozási sebesség is csökken (Viola/Jones, 2001). Viola és Jones alapvetően háromféle *Haar wavelet*et használt, két-téglalap, három-téglalap és négy-téglalap jellemzőket. Az alkalmazásban a detektor Lienhart féle implementációját (Lienhart és Maydt, 2002) használtam, amely tizennégy fajta *wavelet*-et használ („Cascade Classification", 2015).



10. ábra: Kiterjesztett Haar waveletek az OpenCV implementációban

A detektor betanítására az *OpenCV*-ben elérhető *haarcascade_frontalface_alt_tree xml*t használtam, amely korábbi méréseim alapján (Marosvári, 2014) filmes környezetben jól megfelel. Az osztályozó egy csonkolt fa struktúrájú osztályozókból felépített fa struktúra, 20×20px méretű *Haar wavelet*ekkel, *Gentle AdaBoost*tal tanítva.

Az alkalmazás minden talált arcra kiszámolja, hogy a képkocka méretéhez képest mekkora az arc. Ha a képkockán több arcot is talált, akkor az arányszámokat átlagolja, és ez az arányszám lesz a képkocka jellemzője, amely alapján a plánokba sorolást el lehet végezni. Ezzel a módszerrel csak azokat a képkockákat lehet plánokba sorolni, amelyeken látható arc, az arcot nem tartalmazó képek besorolására az algoritmus még nem készült el teljesen.

A rendszer tesztelése

Az elkészült alkalmazás tesztelését egy teljes filmen végeztem el, az *Il deserto rosso* című 1964-es Michelangelo Antonioni filmen. A film hossza 1 óra 52 perc (összesen 168000 képkocka), felbontása 640*352, képsebessége 25 kép/perc. Első lépésben az *ffmpeg* elkészítette a jelenetváltásokat tartalmazó listát: összesen 483 jelenetet különített el. Második lépésben lefuttattam az alkalmazást a teljes filmen. A kapott adatsorokat *RapidMiner*rel beolvastam és elemeztem. A plánok adatsora 561 elemből állt, melyek eloszlása a célváltozók függvényében a következő:

- totál: 21
- kistotál: 70
- amerikai: 32
- szekond: 350
- premier: 20
- ansnitt: 68

Az adatsort 0,632:0,369 arányban osztottam fel tanuló- és teszthalmazba. A tanulóállományba így összesen 355 sor, a teszthalmazba pedig 206 sor került. A kiértékelés során használt mérőszámok a következők:

Fedés: az összes előfordulás hányad részét adta vissza a rendszer.

$$fedés = \frac{TP}{P} = \frac{TP}{(TP+FN)} \quad (1)$$

Pontosság: a rendszer pozitív válaszainak hányad része volt helyes.

$$pontosság = \frac{TP}{(TP+FP)} \quad (2)$$

Accuracy: a rendszer összes válaszának hányad része volt helyes.

$$accuracy = \frac{TP+TN}{(P+N)} = \frac{TP+TN}{(TP+FN+TN+FP)} \quad (3)$$

1. táblázat: Plánfelismerés konfúziós mátrixa RapidMinerrel számítva

	true totál	true kistotál	true szekond	true ansnitt	true amerikai	true premier	osztály pontossága
pred. totál	**1**	3	1	1		1	14,3%
pred. kistotál	1	**8**	1	6	1		47%
pred. szekond	6	8	**121**	15	9	3	74,7%
pred. ansnitt		3	5	**3**	1		25%
pred. amerikai		4			**1**		20%
pred. premier			1			**3**	75%
osztály fedése	12,5%	30,8%	93,8%	12%	8,3%	42,9%	

A plánok felismerésének *accuracy*-je döntési fával: 66,18%. Legnehezebben a szekond, ansnitt és amerikai plánok különböztethetők meg, mert az arcok méretei között nincs nagy különbség. Látható, hogy legjobb eredményt szekond plán esetén ért el a rendszer, a tanulóhalmazban legnagyobb számmal a szekond plán szerepelt, így ennek a felismerése könnyebb is a tanuló rendszernek. Az amerikai és az ansnitt osztályok eredményein azonban látható, hogy valóban könnyen összekeverhetők.

Totál plán esetén a rosszabb fedés az arcdetektor hamis negatív találataiból következik, több vizsgált képkockán sem talált arcot az alkalmazás, így nem tudta felismerni a plánt. Ugyanez a helyzet nagytotál és premier plán esetén is, melyek nem szerepelnek a táblázatban. Nagytotált nem határozott meg a rendszer, ami többnyire helyes eredmény, hiszen nagytotálnál a környezet a hangsúlyos, emberek akár nem is szerepelnek a képen. Szuper közeli plán találat sem volt, ez az arcdetektor működéséből következik: ahhoz, hogy az arc megtalálható legyen, az egész arcnak láthatónak kell lennie, nem csak egyes arcrészeknek. Ezek detektálása szemdetektorral történhet, azonban az ilyen detektorok nem olyan

eredményesek, mint az arcdetektor; sokkal több hamis negatív találatok adnak vissza, ami megnehezíti a helyes plánfelismerést.

Ezekre az esetekre, melyek során nem látható arc a képen, a rendszer jelenleg még nem ad jó eredményt. Az algoritmus pontosítására többféle módszerrel próbálkoztam, SURF pontok térbeli eloszlásának vizsgálatával, éldetektálással, a végső módszer még fejlesztés alatt áll.

További munkám során az előbb említetten kívül szeretném a filmek egyéb képi jellemzőit is felismerni, úgy mint képdinamika, különböző nézőpontok felismerése, illetve a hangsáv elemzésével a filmek hangi jellemzőit is megállapítani.

Irodalomjegyzék

„Cascade Classification." OpenCV 2.4.12.0 documentation, opencv dev team (2015). – 2014.04.27.
http://docs.opencv.org/modules/objdetect/doc/cascade_classification.html

Lienhart, R./Maydt, J.: An Extended Set of Haar-like Features for Rapid Object Detection. *IEEE ICIP* **(1)**, 900–903. 2002.

Marosvári B.: Szereplő azonosítás videó tartalmakban arcdetektálás segítségével. Kézirat. Szakdolgozat. BME-VIK 2014.

„Plán." Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. (2015.11.28.). https://hu.wikipedia.org/wiki/Pl%C3%A1n

„Shot (filmmaking)." Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. (2015.11.28.), https://en.wikipedia.org/wiki/Shot_(filmmaking)

„Shot Types." www.MediaCollege.com, Wavelength Media. (2015.11.28.), http://www.mediacollege.com/video/shots/

Viola, P./Jones, M.: Rapid object detection using a boosted cascade of simple features. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, 12–14. 2001.